JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 2, 273–276 (1961)

On the Functional Equation of Dynamic Programming*

DAVID BLACKWELL

University of California, Berkeley, California

Submitted by Richard Bellman

We imagine a system with a given set S of possible states s. Periodically, say once a day, we observe the current state s of the system, and then choose an action a from a given set A of possible actions. As a joint result of s and a, two things happen: (1) we receive a return $r(s, a)$ (which may be negative) and (2) the system moves to a new state s', selected according to a probability distribution $P_{s,a}$. One of the problems of dynamic programming is to choose a policy which maximizes our expected total return. By a policy π is meant a sequence $\{f_n, n = 1, 2, \ldots\}$ of functions from S to A, with $f_n(s)$ specifying the action to be selected on the nth day, if the system is found to be in state s on that day. We shall denote the expected total return with policy π when the system is initially in state s by $V(s, \pi)$, and the expected total return for the first N days by $V_N(s, \pi)$. We shall assume that

$$\begin{aligned} V(s, \pi) &= r(s, f_1(s)) + \int V(t, \pi_1)\, dP^{(1)}_{s,\pi}(t), \\ V(s, \pi) &= V_N(s, \pi) + \int V(t, \pi_N)\, dP^{(N)}_{s,\pi}(t), \end{aligned} \tag{1}$$

where $P^{(N)}_{s,\pi}$ is the probability distribution of the state of the system at the end of N days, if the system is initially in state s and we use program π, and π_N is the program specified by π beginning with the $N + 1$st day:

$$\pi_N = \{g_n\}, \qquad \text{where} \qquad g_n = f_{N+n}, \quad n = 1, 2, \ldots .$$

* This research was supported by the Information Systems Branch of the Office of Naval Research under Contract Nonr 222(53).

A policy π is called *stable* if f_N is independent of N, i.e., if $\pi_N = \pi$ for all N. For a stable policy π, we obtain from (1):

$$V(s, \pi) \leqslant \sup_a \left| r(s, a) + \int V(t, \pi)\, dP_{s,a} \right|. \tag{2}$$

A policy π^* is called *optimal* if $V(s, \pi^*) \geqslant V(s, \pi)$ for all s, π. If π^* is optimal, we obtain from (1), denoting $V(s, \pi^*)$ by $U(s)$ and by π any policy with $f_1(s) = a$ and $\pi_1 = \pi^*$, the inequality

$$U(s) \geqslant V(s, \pi) = r(s, a) = \int U(t)\, dP_{s,a}(t)$$

so that

$$U(s) \geqslant \sup_a \left[r(s, a) + \int U(t)\, dP_{s,a}(t) \right]. \tag{3}$$

Combining (2) and (3) we obtain that if π^* is a stable optimal policy, its return function $U(s) = V(s, \pi^*)$ satisfies

$$U(s) = \sup_a \left[r(s, a) + \int U(t)\, dP_{s,a}(t) \right]. \tag{4}$$

The problem arises: if we have a stable policy π^* whose return functions satisfies (4), is π^* optimal? The answer would be yes if we know that (a) there is a stable optimal policy and (b) the solution to (4) is unique, and the problem is usually attacked in this way [1]. The purpose of this paper is to show that, in certain circumstances, both (a) and (b) can be avoided: the result is the

THEOREM. *If $\pi^* = \{f, f, \ldots\}$ is a stable policy whose return function satisfies* (4) *and if, for every policy $\pi = \{f_n\}$ and every s,*

$$\sup_N V(s, \pi^{(N)}) \geqslant V(s, \pi) \tag{5}$$

where $\pi^{(N)}$ is the policy which follows π for the first N days and then switches to π^*:

$$\pi^{(N)} = \{g_n\}, \qquad \text{where} \qquad g_n = f_n \quad \text{for} \quad n \leqslant N, \quad g_n = f \quad \text{for} \quad n > N,$$

then π^* is optimal, i.e.,

$$V(s, \pi^*) \geqslant V(s, \pi) \qquad \text{for all } s, \pi.$$

The idea is extremely simple. That the return function for π^* satisfies (4) means that π^* is at least as good as any initial action a, followed by subsequent use of π^*. Examining $\pi^{(N+1)}$ at the beginning of the $N+1$st day, we find that replacing f_{N+1} by f constitutes an improvement, i.e., $\pi^{(N)}$ is an improvement on $\pi^{(N+1)}$. Since $\pi^{(0)} = \pi^*$, π^* is an improvement on every $\pi^{(N)}$. But (5) tells us that, for some N, $\pi^{(N)}$ is almost as good as π, so that π^* is at least as good as π.

Formally, let us say that π' dominates π'' if $V(s, \pi') \geqslant V(s, \pi'')$ for all s. We first note that, for any policy π with $\pi_1 = \pi^*$, i.e., $\pi = (g, f, f, f, \ldots)$ π^* dominates π. For, writing $a_0 = g(s)$

$$V(s, \pi^*) = \sup_a \left[r(s, a) + \int V(t, \pi^*)\, dP_{s,a}(t) \right]$$

$$\geqslant r(s, a_0) + \int V(t, \pi^*)\, dP^{(1)}_{s,a_0}(t) = V(s, \pi).$$

Next, if π' dominates π'', then for any π with $\pi_N = \pi''$, the policy $\bar{\pi}$ obtained by replacing π'' by π' dominates π. For

$$V(s, \bar{\pi}) = V_N(s, \pi) + \int V(t, \pi')\, dP^{(N)}_{s,\pi}(t),$$

$$V(s, \pi) = V_N(s, \pi) + \int V(t, \pi'')\, dP^{(N)}_{s,\pi}(t).$$

Since the first integrand dominates the second, $\bar{\pi}$ dominates π.

It follows from these two facts that $\pi^{(N)}$ dominates $\pi^{(N+1)}$, for

$$\pi^{(N)} = f_1, \ldots, f_N, f, f, f, \ldots,$$

$$\pi^{(NH)} = f_1, \ldots, f_N, f_{N+1}, f, f, \ldots$$

so that $\pi^{(N)}$ is obtained from $\pi^{(N+1)}$ by replacing $(f_{N+1}, f, f, \ldots)$ by the dominating $(f, f, f, \ldots)$.

Thus

$$V(s, \pi^*) \geqslant V(s, \pi^{(N)}) \qquad \text{for all } N, s.$$

Invoking hypothesis (5) proves the theorem.

We note finally that (5) will not always hold. For example, consider a system with two states, 0 and 1, and two actions, 0 and 1. In state 0, no matter which action we choose, the new state is 0 and our income is 0. In state 1, action 1 keeps the system in state 1 and gives us an income of 0, while action 0 moves the system to state 0 and costs us 1 unit.

If π is the stable program with $f(0) = 0$, $f(1) = 1$,

$$V(\pi, s) = 0 \qquad \text{for} \qquad s = 0, 1.$$

If π^* is the stable program with $f(0) = f(1) = 0$, we have

$$V(\pi', 0) = 0, \qquad V(\pi', s) = -1.$$

Equation (4) becomes

$$U(0) = U(0),$$

$$U(1) = \max [U(1), -1 + U(0)],$$

that is,

$$U(1) \geqslant U(0) - 1.$$

Both $V(s, \pi)$ and $V(s, \pi^*)$ satisfy this inequality, so that (5) must be violated. Indeed if $\pi^{(N)}$ is the policy which follows π for the first N days and then switches to π^*, we have

$$V(1, \pi^{(N)}) = -1 \qquad \text{for all } N,$$

while $V(1, \pi) = 0$.

Reference

1. Bellman, Richard, "Dynamic Programming." Princeton Univ. Press, Princeton, New Jersey, 1957.